

Notities en Commentaren

MAAR RAAKT WEL ZIJN EIGEN VOET

W.J. van der Linden¹ en M.A. Zwarts²

¹Universiteit Twente en ²Inspectie van het Onderwijs

Het blijft moeilijk om in de betogen van Treffers een consistente lijn te ontdekken. Wat moeten we denken van een auteur die een paar jaar geleden de representativiteit van de PPON itempool nog de hemel in prees (Treffers, 1988; zie ook onze vorige reactie), nu uit twee merkwaardige stellingen het tegendeel concludeert, maar vervolgens wel in een adem doorgaat en verheugd meedeelt dat de resultaten uit PPON-2 overtuigend aantonen dat de rekenmethoden die zich op het werk van het Freudenthal Instituut baseren het beter doen (Treffers, 1996b)? Wat voor waarde heeft deze conclusie als het PPON-materiaal niet representatief en dus niet inhoudsvalide is? In zijn eerste notitie schoot Treffers mis, maar thans lijkt het er verdraaid veel op dat hij in zijn eigen voet heeft geschoten. Gelukkig maar dat de CEB bij zijn oordeelsvorming over de kwaliteit van het rekenonderwijs wel aandachtig kennis heeft genomen van de commentaren van Treffers, maar er niet blind op heeft gevaren.

Halve waarheden en hele suggesties

In de tweede notitie van Treffers staan nogal wat halve waarheden en hele suggesties. We laten deze puntsgewijs de revue passeren en nemen daarbij de volgorde in acht die Treffers in deze notitie heeft aangehouden.

1. In stelling 1 wordt betoogd dat als de CEB-normen op het beste rekenland ter wereld - volgens Treffers was dat toch Nederland? - toegepast worden, er geen voldoende uit de bus zal komen. Dit is een experiment dat de CEB nooit voor haar rekening zou willen nemen. In evaluatie-onderzoek is het absoluut onjuist om eisen aan de opbrengsten van een vak te stellen zonder rekening te houden met de condities waarover de onderwijsgegenden zelf geen controle hebben. De CEB heeft dit reeds gesteld in zijn plan voor de evaluatie van het basisonderwijs en als voorbeelden van zulke condities genoemd het landelijk onderwijsbeleid en de daarop gebaseerde uitgaven, de materiële voorzieningen (o.a. leer- en hulpmiddelen en de mogelijkheden van onderwijs-op-maat), de kenmerken van de schoolbevolking (o.a. etniciteit, gezinssituatie en schoolverzuim) en de kenmerken van de leraarpopulatie (o.a. opleidingsniveau en arbeidssatisfactie) (Commissie Evaluatie Basisonderwijs, 1992, pag. 13). Voor de opbrengsten van een individueel vak komt daar nog bij het beslag dat de andere schoolvakken op de beschikbare tijd leggen. De normen die de CEB heeft aangelegd, zijn gebaseerd op een zo compleet mogelijke empirische beschrijving van de toestand van het basisonderwijs. In andere landen gelden andere condities en is de beschikbare onderwijstijd anders verdeeld. Als de werkwijze van de CEB elders herhaald zou moeten worden, zouden er dus ongetwijfeld andere normen uit rollen. Iedere serieuze evaluator zal hetzelfde principe volgen. Als de *Consumentenbond* stereoversterkers beoordeelt, houdt ze ook rekening met de prijs van de aanbevolen apparaten.

2. In stelling 2 wordt beweerd dat het mogelijk is om toetsen te ontwikkelen die psychometrisch identiek zijn maar verschillende scoreverdelingen opleveren. Hierover kunnen we kort zijn. Kennisname van twee bladzijden in een leerboek over psychometrie is voldoende om te weten dat verschillende test- en itemparameters verschillende scoreverdelingen impliceren. Deze stelling is klinkklare onzin.

3. De bewijsvoering voor beide stellingen moet het twee keer doen met een beroep op anonieme autoriteiten die Treffers zouden bijvallen. Zo adstrueert men niet in een wetenschappelijke discussie. Wie zijn al die vakdeskundigen die het met stelling 1 eens zijn? Onder wiskundigen wordt bepaald verschillend gedacht over de kwaliteit van het rekenonderwijs. Direct

Adres: Universiteit Twente, Faculteit Toegepaste Onderwijskunde, Postbus 217, 7500 AE Enschede.

na het vaker gememoreerde interview met Treffers in het *NRC Handelsblad* kon men in dezelfde courant een ingezonden stuk lezen van de wiskundigen Barendregt (KUN) en Ruttkay (VU), die zich zeer bezorgd uitlieten over het Nederlandse rekenonderwijs en kritiek hadden op de aanpak die het Freudenthal Instituut voorstaat. Opnieuw in dezelfde courant werd op 6 september j.l. een artikel afgedrukt van de hand van H.K. Kuiken, hoofdredacteur van het *Journal of Engineering Mathematics*, waarin deze betoogde dat hij op grond van zijn ervaringen met het universitaire toelatingsexamen in China - het land dus waar volgens Treffers de Nederlandse aanpak wereldberoemd is - tot de conclusie komt dat vrijwel geen Nederlandse kandidaat het onderdeel met goed gevolg zou doorstaan. Ook wijst hij erop dat in de wiskunde-olympiade Nederland de laatste jaren onveranderlijk tussen de dertigste en de veertigste plaats figureert. En wie zijn die Cito-vakdeskundigen die de merkwaardige bewering in stelling 2 voor hun rekening durven nemen? Ze bestaan niet of zouden op staande voet ontslagen moeten worden.

4. De waarheid van de eerste stelling meent Treffers verder aan te kunnen tonen door een "vergelijking van overeenkomstige items in verschillende nationale onderzoeken" en door "de resultaten van het tweede en het derde internationale onderzoek voor rekenen-wiskunde van de IEA te analyseren". Welke nationale onderzoeken heeft Treffers precies geraadpleegd? Buiten Australië, de V.S. en Nederland zijn er amper landen waar ooit een nationale peiling van de opbrengsten van het basisonderwijs op basis van een representatieve steekproeftrekking is uitgevoerd. Of gaat het hier om buitenlandse herhalingen van het in onze vorige reactie op Treffers gememoreerde onderzoek op "enkele scholen in de buurt van Utrecht die zeker representatief zijn voor heel Nederland"? Bovendien weten alle toetsconstructeurs dat kleine verschillen in de bewoording of de vraagvorm van een item tot even onvoorspelbare als drastische verschuivingen in de moeilijkheid ervan kunnen leiden. Zie bijvoorbeeld de opmerking terzake van Wijnsstra (1995). Scoreverdelingen op één verzameling items voorspellen uit verdelingen op een andere verzameling kan niet op grond van didactische analyse van de inhoud van opgaven, maar alleen met behulp van steekproeftechnieken en statistische analyses die bekend staan onder de naam testequivalering. Voor een recent overzicht van degelijke methoden verwijzen we onze opponent naar Kolen en Brennan (1995). Wat de IEA-toetsen betreft, deze zijn veel te beperkt om uitspraken te kunnen doen op het niveau van de kerndoeldomeinen voor rekenen die de CEB in zijn oordelen heeft aangehouden. Trouwens, het TIMSS-project, waarbinnen de derde IEA-peiling van het rekenonderwijs wordt uitgevoerd, heeft zijn toetsen nog niet zo lang geleden afgenomen en zit nu midden in de voorbereidingen van zijn data-analyses. Resultaten van internationale vergelijkingen zijn op dit moment nog niet beschikbaar.

5. Het gebrek aan representativiteit van de itempool uit PPON meent Treffers verder aan te kunnen tonen met "elementaire vakinhoudelijke middelen". Het wondermiddel blijkt een taxonomie voor de leerstof over breuken te zijn die ontstaat door vier verschillende indelingscriteria met elkaar te combineren. Maar wat kun je nu in vredesnaam met een taxonomie bewijzen? Met het toevoegen van indelingscriteria aan taxonomieën kun je blijven doorgaan en iedere didacticus heeft waarschijnlijk zijn eigen voorkeur voor een taxonomie. Voor de vraag of de verzameling PPON-opgaven een representatieve steekproef uit het leerstofdomein vormt, zijn deze indelingen statistisch gezien van geen enkel belang. Het enige wat ze in feite mogelijk maken, is de trekking van een gestratificeerde in plaats van een enkelvoudig aselechte steekproef items. Ironische genoeg leert de steekproeftheorie ons dat in het eerste geval het aantal items in de steekproef juist kleiner kan zijn om de aantal-goedscore op het domein met een gelijke precisie te schatten (zie bijvoorbeeld Bhattacharyya & Johnson, 1977, hoofdstuk 16). Het belangrijkste is evenwel het volgende. Wat voor het onderwerp telt, zijn niet de theoretische discussies over de juiste taxonomie, maar het *empirische antwoord op de vraag met hoeveel dimensies de leerstof en de itemverzameling beschreven moet worden*. De in PPON gebruikte methodologie van een op een IRT-model gebaseerde schalingsmethode is een effectief antwoord op deze vraag. Als het aantal dimensies bekend is, is het enige waar de verdeling van de aantal-goedscores nog van afhangt de verdeling van de itemparameters over deze schalen. Door middel van robuustheidsanalyse heeft de CEB laten zien dat haar oordelen amper gevoelig waren voor verschuivingen in deze parameterverdelingen (Van der Linden & Zwarts, 1995a). Dit is op-

nieuw door ons betoogd in onze reactie op Wijnstra (Van der Linden & Zwarts, 1995b) en tenslotte nog een keer in onze vorige reactie op Treffers (Van der Linden & Zwarts, 1996). We vragen ons zo langzamerhand oprecht af of Treffers onze stukken wel leest.

6. In de volgende alinea meent Treffers ons te moeten wijzen op de moeilijkheidsgraad van de opgaven in de PPON-pool. De suggestie dat deze ons niet bekend zouden zijn, is wel heel pedant. Zoals al het overige materiaal, beschikte de CEB ook over deze informatie. Datafiles met de itemparameters werden ons door het Cito welwillend ter beschikking gesteld en zijn uitvoerig gebruikt in ons rekenwerk. Het meest krasse is nog wel dat Treffers weet dat één van ons ten tijde van de rekenpeiling als Cito-medewerker op PPON werkte. Nee hoor, de CEB heeft deze informatie heus niet vergeten toen zij haar beoordelingen opstelde.

7. In zijn analyse van de vraag of er sinds 1987 veel in het rekenonderwijs is gewijzigd, heeft onze opponent alleen oog voor het didactische karakter van de gebruikte leermethoden—en dan nog slechts voor de vraag of ze zich baseren op de ideeën van het Freudenthal Instituut. Kennelijk verwacht hij aardverschuivingen in de opbrengsten van het onderwijs ten gevolge van een verandering van leerboek. Zo gaat het er in het onderwijs niet aan toe. Er zijn factoren die veel krachtiger zijn. De betekenis van leerboeken zinkt zelfs volledig in het niet als, zoals in het PPON-onderzoek is gebleken, een aanzienlijk aantal scholen amper toekomt aan de behandeling van sommige onderwerpen. Voor andere voorbeelden van effectiviteitsbepalende factoren verwijzen we naar de uitkomsten van een meta-analyse in Fraser (1989). De CEB stond voor de taak om het basisonderwijs in al zijn aspecten te evalueren. Onze stelling blijft dat er in de betreffende 5 jaar, met uitzondering van de toename van het gebruik van realistische methoden, in het geheel van het rekenonderwijs niet veel is veranderd. Belangrijk is dat de lezer opmerkt dat het kwantitatieve overzicht dat Treffers schijnbaar ter ondersteuning van zijn bewering geeft, slechts slaat op de verschillen in opbrengsten tussen traditionele en realistische rekenmethoden en dus *niet* op de vergelijkingen van de opbrengsten tussen 1987-1992. (De tweede peiling vond plaats in 1992 en niet in 1994 zoals Treffers lijkt te willen suggereren.) Deze verschillen waren al in 1987 bekend en volgen het vaste patroon uit het leerboekenonderzoek dat methoden het beter doen op de onderdelen die hun auteurs belangrijk vinden en slechter op de onderdelen die ze minder belangrijk vinden. In tegenstelling tot Treffers heeft de CEB, direct na het beschikbaar komen van de laatste gegevens, de consequenties van het verschil in opbrengsten tussen 1987-1992 voor haar beoordelingen van het rekenonderwijs wel geanalyseerd (Zwarts, 1994). De uitkomst was dat op het niveau van de 27 individuele rekenschalen er 2 positiever (gehele getallen en hoofdrekenen met vermenigvuldigen en delen) en 2 negatiever (cijferend vermenigvuldigen en berekenen van het aantal maateenheden) beoordeeld zouden zijn geworden. Worden de oordelen evenwel geaggregeerd op het niveau van de kerndoel domeinen, en dat is het rapportageniveau dat de CEB in zijn eindrapport aangehouden heeft, dan vallen deze verschuivingen tegen elkaar weg en blijft het oordeel onveranderd.

8. Tenslotte dan de kwestie van de vier dansopgaven die Treffers in zijn vorige notitie formuleerde. De lezer moet ze nog maar eens nalezen en voor zichzelf de vraag beantwoorden of de beheersing van de laatste twee wel tot het basisonderwijs gerekend mag worden. Wij blijven apert van mening dat dit niet het geval mag zijn. Deze opgaven verschillen veel te veel in niveau van de PPON-opgaven, die Treffers gemiddeld al te moeilijk voor het basisonderwijs vond. Opgaven die een algemeen bewijs vragen ("Toon aan dat ...") mogen niet voorgelegd worden aan leerlingen die nog enkel geen idee over methoden van bewijsvoering hebben. Dat onder de vele manieren waarop een opgave opgelost kan worden er één is die eenvoudig blijkt te zijn, betekent nog niet dat de opgave voor de leerlingen van de basisschool eenvoudig is. We gaan dan nog voorbij aan het feit dat de redactie van de gemeenschappelijk stam van deze vier opgaven niet specifiek genoeg is voor leerlingen van de basisschool en dat reeksen opgaven waarvan het antwoord op de volgende vraag afhangt van succes op een voorafgaande vraag meettechnisch gezien uit de boze zijn. Dergelijke zaken tellen mee als de moeilijkheid van opgaven empirisch wordt vastgesteld. We zien hier opnieuw een fraaie demonstratie van het feit dat als didactici lang met een onderdeel van de leerstof bezig zijn, ze de neiging krijgen om de moeilijkheid ervan te gaan onderschatten.

SLOTWOORD

Wat ons betreft zijn we aan het einde gekomen van onze discussie met Treffers. Het is moeilijk discussiëren met een opponent die slechts oog heeft voor het didactische aspect van het rekenonderwijs, die zijn theorieën baseert op informele analyses van losse opgaven, zich voor evidentie beroept op meningen van gelijkgezinden en weinig interesse toont voor de statistische methodologie van peilingsonderzoek. Bovendien reageert hij in zijn tweede notitie nergens op de eerder door ons aangevoerde argumenten, zodat een echte discussie niet ontstaat. Opnieuw lijkt het er verdraaid veel op alsof hij in het wildeweg schiet in de veronderstelling dat rook de beste manier is om vuur te suggereren. Dat hij zijn eigen voeten raakt en vervolgens in zijn betoog over de betekenis van het realistische rekenonderwijs eigenlijk geen been meer heeft om op te staan, komt niet voor onze rekening. Voor een eerlijk duel zijn we altijd in, maar van losvliegende kogels blijven we graag verschoond.

LITERATUUR

- Bhattacharyya, G.K., & Johnson, R.A. (1977). *Statistical concepts and methods*. New York City, NY: Wiley.
- Commissie Evaluatie Basisonderwijs (1992). *Plan voor de evaluatie van het basisonderwijs*. De Meern: Inspectie van het Onderwijs.
- Fraser, B.J. (1989) Research syntheses on school and instructional effectiveness. *International Journal of Educational Research*, 13, 707-720.
- Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and practices*. New York City, NY: Springer-Verlag.
- Van der Linden, W.J. & Zwarts, M.A. (1995a). Robustness of judgments in evaluation research. *Tijdschrift voor Onderwijsresearch*, 20, 13-27.
- Van der Linden, W.J. & Zwarts, M.A. (1995b). De opbrengsten van het basisonderwijs: een repliek. *Tijdschrift voor Onderwijsresearch*, 20, 34-41.
- Treffers, A. (1988). Over de merkbare invloed van onderwijsmethoden op leerprestaties. In J.M. Wijnstra (red). *Balans van het rekenonderwijs in de basisschool*. Arnhem: Cito.
- Treffers, A. (1996a). Rekeninhoudelijke kritiek op CEB-repliek. *Tijdschrift voor Onderwijsresearch*, 21,
- Treffers, A. (1996b). Rekeninhoudelijke kritiek op CEB-repliek (2).
- Wijnstra, J.M. (1995). De opbrengsten van het basisonderwijs volgens de CEB: Enkele kanttekeningen bij de gevolgde normeringsprocedure. *Tijdschrift voor Onderwijsresearch*, 20, 28-33.
- Zwarts, M.A. (1994). *Beoordeling gegevens nieuwe rekenpeiling* (intern memorandum). De Meern: Inspectie van het Onderwijs.

NOOT VAN DE REDACTIE

In dit commentaar reageren de auteurs op een niet gepubliceerde dupliek van Treffers. Nadat zowel de dupliek van Treffers als dit weerwoord daarop van Van der Linden en Zwarts door de redactie waren geaccepteerd, trok de heer Treffers zijn dupliek terug. De redactie betreurt de gang van zaken. Op verzoek van de heren Van der Linden en Zwarts wordt hun tweede commentaar toch gepubliceerd. Geïnteresseerde lezers kunnen de tekst van de dupliek van Treffers bij de auteur opvragen. Zijn adres is: Prof.dr. A. Treffers, Freudenthal Instituut, Tiberdreef 4, 3561 GG Utrecht.